A Matrix Minimization Problem Involving Ranks

J. Ponstein
Econometrics Institute
University of Groningen
P. O. Box 800
9700 AV Groningen, The Netherlands

Submitted by Richard A. Brualdi

ABSTRACT

The problem of finding a matrix whose rank is limited from above such that the norm of the difference of this matrix and a given matrix is minimal, is approached from the point of view of the theory of generalized gradients and normal cones. The matrices involved are assumed to be finite, although some results can be generalized without further assumptions to the case where the number of rows and columns is countably infinite. Two norms will be considered, i.e. the l_2-induced norm and the Frobenius norm, with an emphasis on the latter. Special attention is paid to what will be called reversely circulant matrices, which are finite Hankel matrices of a special nature.

1. INTRODUCTION

The problem we are going to consider is, given a finite, not necessarily square, matrix Q^0, and $k < \operatorname{rank} Q^0$, to find a matrix Q of the same size as Q^0, such that the norm of $Q - Q^0$ is minimized under the condition that $\operatorname{rank} Q \leqslant k$:

$$\inf_{Q} \{ |Q - Q^0| : \operatorname{rank} Q \leqslant k \}, \qquad k < \operatorname{rank} Q^0.$$

Here it may happen that the rank of the minimizing Q is less than k (see Example 6 at the very end of the paper). Hence requiring that $\operatorname{rank} Q = k$ would lead to another (probably easier) problem.

0024-3795/87/$3.50

Some of the results are also valid for matrices with a countably infinite number of rows and columns.

Special attention is devoted to matrices Q^0 and Q of the form

$$\begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ a_2 & a_3 & \cdots & a_1 \\ \cdots & \cdots & \cdots & \cdots \\ a_n & a_1 & \cdots & a_{n-1} \end{pmatrix},$$

which we call *reversely circulant*. Recall that a *circulant* matrix has the form

$$\begin{pmatrix} a_1 & a_2 & \vdots & a_n \\ a_n & a_1 & \cdots & a_{n-1} \\ \cdots & \cdots & \cdots & \cdots \\ a_2 & a_3 & \cdots & a_1 \end{pmatrix}.$$

We will approach this minimization problem by using *generalized gradients* and *normal cones*, introduced by Clarke [1] and investigated by him, Rockafellar [2,3], and others. Define f and C by

$$f(Q) = |Q - Q^0|, \qquad C = \{Q : \operatorname{rank} Q \leqslant k\}.$$

Further, let $\partial f(Q)$ be the set of all *generalized gradients of f at Q*, and let $N_C(Q)$ be the *cone of normals* (or *normal cone*) *of C at Q*. Then, since C is closed and f is convex and Lipschitz in a neighborhood of any Q, it follows from Rockafellar [2, Theorem 5K] or Clarke [1, Corollary to Proposition 2.4.3] that for any optimal Q,

$$0 \in \partial f(Q) + N_C(Q).$$

Since in most cases C is not convex, this is only a necessary condition for optimality. In order to find a global minimum it will therefore be necessary to find first all local minima; see below for details and examples.

The usual approaches are quite different from the one we will follow. One is by applying results about the the *singular values* of Q^0; see e.g. Stewart [4]. In particular, if the number of rows and the number of columns are countably infinite, and if $|Q - Q^0|$ is the l_2-induced norm, that is, if

$$|Q - Q^0|^2 = \sup_x \left\{ \left| (Q - Q^0)x \right|^2 : |x| \leqslant 1 \right\} = \lambda_{\max}\left((Q - Q^0)^T (Q - Q^0) \right),$$

where both $|(Q - Q^0)x|$ and $|x|$ are l_2-norms, and Q and Q^0 must, of course, belong to the space of matrices whose l_2-induced norm is finite, then it is known that if the singular values of Q^0 are, in descending order, $\sigma_1, \sigma_2, \ldots$, the infimum is a minimum and is equal to σ_{k+1}. Moreover, the minimum is assumed for a Hankel matrix if Q^0 itself is a Hankel matrix (a Hankel matrix is characterized by the fact that each element depends on only the sum of the row index and the column index). If the matrices are finite, the minimum is still equal to σ_{k+1}, but is not necessarily assumed for some Hankel matrix (see a counterexample in Section 6, below). Requiring that Q^0 be reversely circulant, it is an open question whether or not the minimum is assumed for some reversely circulant matrix.

Obviously, a disadvantage of working with singular values is that this seems difficult to combine with structural properties of the matrices involved, such as being Hankel or reversely circulant. Anyhow, it seems that characterizing all optimal solutions (possessing certain structural properties), if the matrices are of infinite size and if the l_2-induced norm is adopted, is a difficult problem. Perhaps, the approach followed in the present paper might be helpful in this respect, even when the matrix elements, which we assume to be real numbers, are replaced by matrices, as in *systems theory* (see e.g. Glover [5]).

Speaking about norms, it is appropriate to remark that although the l_2-induced norm is convenient for the approach reviewed above, it is not for the approach we want to follow, simply because then $\partial f(Q)$ usually is not a singleton. If instead we take the *Frobenius norm*, defined by

$$|Q - Q^0|^2 = \sum_{i,j} \left(q_{ij} - q_{ij}^0\right)^2,$$

where q_{ij} and q_{ij}^0 are the ij-elements of Q and Q^0, respectively, then $\partial f(Q)$ is a singleton if $Q \neq Q^0$, and this requirement is satisfied, as rank $Q <$ rank Q^0. As a result, the condition $0 \in \partial f(Q) + N_C(Q)$ becomes tighter. Because of this and because the choice of the norm sometimes is not too critical, we will put an emphasis on the use of the Frobenius norm. An additional advantage is that f then is ordinarily differentiable at Q, so that the main problem that remains is to determine $N_C(Q)$.

No matter which norm is selected, we must (assuming that Q^0 and Q are m by n) topologize the space of all m by n matrices. This is done via the norm itself, so that the result is a normed space of dimension mn. The dual of this space is the set of all linear functionals

$$Z: Q \mapsto (Z, Q),$$

which can be represented by m by n matrices Z such that (with a slight abuse of notation)

$$(Z, Q) = \sum_{i,j} z_{ij} q_{ij} = \operatorname{trace}(Z^T Q).$$

Our program will be as follows. First we will compute $N_C(Q)$, the normal cone of C at Q, where C is the feasible region of our optimization problem. This is done by first computing $T_C(Q)$, which is the *cone of tangents of C at Q*, as introduced by Clarke [1], and then applying polarization to $T_C(Q)$.

The next step is to compute $\partial f(Q)$ for the l_2-induced norm as well as for the Frobenius norm.

Then we will see what can be deduced from the inclusion $0 \in \partial f(Q) + N_C(Q)$, in particular if the norm is the Frobenius norm and if the matrices involved are reversely circulant. In the latter case a complete computational scheme for finding $\inf_Q\{|Q - Q^0| : \operatorname{rank} Q \leqslant k\}$ will be given, by considering separately Q's with $\operatorname{rank} Q = p$ for $p = k, k-1, \ldots$.

One might wonder why it is necessary to use Clarke's cone of tangents $T_C(Q)$ rather than Bouligand's *contingent cone of C at Q*, which we indicate by $K_C(Q)$. The latter is the set of all matrices V such that there exists a sequence $t_i \to 0$, $t_i > 0$, as well as a sequence $V_i \to V$ such that $Q + t_i V_i \in C$ for all i [a definition of $T_C(Q)$ is given at the beginning of the next section]. If we could work with $K_C(Q)$ instead of $T_C(Q)$, this should certainly be preferred, as $K_C(Q)$ is easier to compute and to analyse than is $T_C(Q)$. However, whereas $T_C(Q)$ is always closed and convex, $K_C(Q)$, although closed, need not be convex, and probably this is the reason why there is no satisfactory theory leading to inclusions like

$$0 \in \partial f(Q) + N_C(Q),$$

with $N_C(Q)$ the polar cone of $K_C(Q)$.

To be more specific, consider the case where the matrices involved are 3 by 3, $k = 2$, and

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Then

$$T_C(Q) = \left\{ \begin{pmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & 0 & 0 \\ v_{31} & 0 & 0 \end{pmatrix} \right\}$$

and

$$K_C(Q) = \left\{ \begin{pmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{pmatrix} : \det \begin{pmatrix} v_{22} & v_{23} \\ v_{32} & v_{33} \end{pmatrix} = 0 \right\},$$

so that $T_C(Q)$ is different from $K_C(Q)$. Since we always have that $K_C(Q) \supset T_C(Q)$, this means that this inclusion is strict, and that $K_C(Q)$ might be too large, so that its polar might be too small. Moreover, this $K_C(Q)$ is not convex, as is easily verified.

In this example $\operatorname{rank} Q < k$, so let us now consider the case where $\operatorname{rank} Q = k$. Then it can be shown that $T_C(Q) = K_C(Q)$. So, if we required that $\operatorname{rank} Q = k$ rather than $\operatorname{rank} Q \leqslant k$, then using Bouligand's contingent cone would do, but, as pointed out at the beginning of this section, this would lead to another problem.

2. COMPUTING $N_C(Q)$

Recall that, with k given,

$$C = \{ Q : \operatorname{rank} Q \leqslant k \}.$$

By definition,

$$N_C(Q) = \{ Z : (Z, V) \leqslant 0 \text{ for all } V \in T_C(Q) \},$$

where $T_C(Q)$, again by definition, is the set of all V such that for all sequences $Q_1, Q_2, \ldots$ and $t_1, t_2, \ldots$ satisfying $Q_i \in C$, $t_i > 0$, $Q_i \to Q$, $t_i \to 0$, we can find a sequence $V_1, V_2, \ldots$ satisfying $Q_i + t_i V_i \in C$ and $V_i \to V$.

THEOREM 1. *If* $\operatorname{rank} Q \leqslant k$, *with* Q m *by* n, *and if* $k < \min(m, n)$, *then* $V \in T_C(Q)$ *if and only if* $V(\ker Q) \subset \operatorname{im} Q$. *Hence for* Q *fixed, increasing* k, *but so that* $k < \min(m, n)$ *remains true, does not change* $T_C(Q)$.[1]

Proof.

(A) Let $V \in T_C(Q)$, and let $t_i \downarrow 0$. Since for nonsingular A and B of appropriate sizes we have $V(\ker Q) \subset \operatorname{im} Q$ if and only if $AVB(\ker AQB) \subset$

[1] The last observation is due to the referee.

im AQB, we may assume that Q takes the form

$$Q = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}$$

with A p by p, $p \leqslant k$, and nonsingular. Let $s_i = t_i^{1/2}$, and let

$$Q_i = \begin{pmatrix} A & 0 & 0 \\ 0 & s_i I & 0 \\ 0 & s_i L & 0 \end{pmatrix},$$

where I is the $k-p$ by $k-p$ identity matrix, and L is a fixed $n-k$ by $k-p$ matrix. So $\operatorname{rank} Q_i = k$, $Q_i \to Q$. Hence $V_i \to V$ must exist such that $\operatorname{rank}(Q_i + s_i^2 V_i) \leqslant k$. Partition V_i into submatrices V_{qr}^i, $q, r = 1,2,3$, and let $V_{qr} = \lim V_{qr}^i$. If i is large enough,

$$\begin{pmatrix} A + s_i^2 V_{11}^i & s_i^2 V_{12}^i \\ s_i^2 V_{21}^i & s_i I + s_i^2 V_{22} \end{pmatrix}$$

is nonsingular, so that $\operatorname{rank}(Q_i + s_i^2 V_i) \geqslant k$. But $\operatorname{rank}(Q_i + s_i^2 V_i) \leqslant k$; hence $\operatorname{rank}(Q_i + s_i^2 V_i) = k$, and matrices K_1^i and K_2^i must exist such that

$$s_i^2 V_{31}^i = K_1^i\big(A + s_i^2 V_{11}^i\big) + K_2^i s_i^2 V_{21}^i,$$

$$s_i L + s_i^2 V_{32}^i = K_1^i s_i^2 V_{12}^i + K_2^i\big(s_i I + s_i^2 V_{22}^i\big),$$

$$s_i^2 V_{33}^i = K_1^i s_i^2 V_{13}^i + K_2^i s_i^2 V_{23}^i.$$

As A is nonsingular, it follows from the first two of these equations that K_2^i tends to L and that K_1^i tends to 0, so that the last equation implies that $V_{33} = LV_{23}$. But L is arbitrary; hence $V_{33} = 0$ and $V_{23} = 0$. By symmetry we may interchange the indices 2 and 3, so that also $V_{22} = 0$ and $V_{32} = 0$, and hence $V(\ker Q) \subset \operatorname{im} Q$. Notice that this conclusion is only correct if V_{33} (or V_{22}) is nonempty, which is true because $k < \min(m, n)$.

(A′) A coordinate-free proof of this part of the proof has been given by Nieuwenhuis [6]. This leads to the following alternative. Let $X + Y + Z = R^n$ be such that $\dim X = p$, $\dim Y = k - p$, $\dim Z = n - k$, $Q(Y + Z) = 0$, and $QX = \operatorname{im} Q$. Define Q_i as follows: $Q_i x = Qx$ if $x \in X$, $Q_i y = s_i \tilde{y}$ if $y \in Y$ for some $\tilde{y} \in \tilde{Y}$, a fixed $k - p$ dimensional subspace of $\operatorname{im}^c Q$ (the complement of

$\operatorname{im} Q$), and $Q_i Z = 0$. Then $\operatorname{rank} Q_i = k$ and $Q_i \to Q$, so that $V_i \to V$ must exist such that $\operatorname{rank}(Q_i + s_i^2 V_i) \leqslant k$. But $(Q_i + s_i^2 V_i)X = \operatorname{im} Q + s_i^2 V_i X$ and $(Q_i + s_i^2 V_i)Y = \tilde{Y} + s_i^2 V_i \tilde{Y}$, so that if i is large enough $(Q_i + s_i^2 V_i)(X + Y)$ is at least $p + (k - p) = k$ dimensional. Hence $\operatorname{rank}(Q_i + s_i^2 V_i) = k$, and given any $z \in Z$, $V_i z = (Q_i + s_i^2 V_i) s_i^{-2} z = Q x_i + \tilde{y}_i$ for suitable $x_i \in X$ and $\tilde{y}_i \in \tilde{Y}$. Hence $Vz = Qx + \tilde{y}$ for suitable $x \in X$, $\tilde{y} \in \tilde{Y}$. But $\operatorname{im}^c Q$ is $m - p$ dimensional; hence, since $k < m$, and thus $k - p < m - p$, we can find $\tilde{y}' \in \tilde{Y}$, independent of $\tilde{y}$, such that for some $\alpha \in R$, $Vz = Qx + \alpha \tilde{y}'$, so that $\tilde{y} = \alpha \tilde{y}'$, and hence $\tilde{y} = 0$ and $Vz = Qx$. But since $k < n$, Z is at least 1 dimensional, so that we can choose Y and Z such that any given element of $\ker Q$ is in Z, and it follows that $V(\ker Q) \subset QX = \operatorname{im} Q$.

(B) Conversely, let $V(\ker Q) \subset \operatorname{im} Q$, and let $t_i \downarrow 0$, $Q_i \to Q$, $\operatorname{rank} Q_i \leqslant k$ be given. Again we may assume that Q takes the form

$$Q = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix},$$

with A as before. We must find $W_i \to 0$ such that $\operatorname{rank}(Q_i + t_i V + t_i W_i) \leqslant k$. Consider a subsequence such that $\operatorname{rank} Q_i = k'$ is constant and there is a constant set of k' independent rows of Q_i, including those of $(A\ 0)$. That is, let Q_i, up to a permutation of rows and columns, take the form

$$Q_i = \begin{pmatrix} Q_{11}^i & Q_{12}^i & Q_{13}^i \\ Q_{21}^i & Q_{22}^i & Q_{23}^i \\ Q_{31}^i & Q_{32}^i & Q_{33}^i \end{pmatrix},$$

where

$$\operatorname{rank} \begin{pmatrix} Q_{11}^i & Q_{12}^i \\ Q_{21}^i & Q_{22}^i \end{pmatrix} = k'$$

for a subsequence. It follows that L_1^i and L_2^i must exist such that

$$(Q_{31}^i, Q_{32}^i, Q_{33}^i) = L_1^i (Q_{11}^i, Q_{12}^i, Q_{13}^i) + L_2^i (Q_{21}^i, Q_{22}^i, Q_{23}^i).$$

This may, however, lead to divergent L_2^i and L_1^i. In order to avoid this, check for each row r_3 of $(Q_{31}^i, Q_{32}^i, Q_{33}^i)$ in succession whether or not certain elements of the row of L_2^i corresponding to r_3 diverge to either $+\infty$ or $-\infty$ for a subsequence. If so, select the fastest divergent element, and interchange

the row of $(Q_{21}^i, Q_{22}^i, Q_{23}^i)$ that corresponds to this element with r_3. In this way we can make sure that L_2^i converges for a subsequence. Since $Q_i \to Q$, so that $Q_{11}^i \to A$, it follows that $L_1^i \to 0$. Since $V(\ker Q) \subset \operatorname{im} Q$, it further follows that

$$V = \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & 0 & 0 \\ V_{31} & 0 & 0 \end{pmatrix}.$$

Let

$$W_i = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & W_{32}^i & W_{33}^i \end{pmatrix}.$$

It easily follows that $\operatorname{rank}(Q_i + t_i V + t_i W_i)$ is equal to the rank of

$$\begin{pmatrix} Q_{11}^i + t_i V_{11} & Q_{12}^i + t_i V_{12} & Q_{13}^i + t_i V_{13} \\ Q_{21}^i + t_i V_{21} & Q_{22}^i & Q_{23}^i \\ V_{31} - L_1^i V_{11} - L_2^i V_{21} & W_{32}^i - L_1^i V_{12} & W_{33}^i - L_1^i V_{13} \end{pmatrix}.$$

Try to find K^i such that the last "row" of this matrix is K^i times the first one. Then its rank will be at most k. Hence put

$$V_{31} - L_1^i V_{11} - L_2^i V_{21} = K^i\left(Q_{11}^i + t_i V_{11}\right),$$

$$W_{32}^i - L_1^i V_{12} = K^i\left(Q_{12}^i + t_i V_{12}\right),$$

$$W_{33}^i - L_1^i V_{13} = K^i\left(Q_{13}^i + t_i V_{13}\right).$$

Since Q_{11}^i tends to the nonsingular A, we can solve for K^i from the first of these equations, and since $L_1^i \to 0$, it follows from the remaining two equations that $W_{32}^i \to 0$, $W_{33}^i \to 0$, which means that we have found the required W_i, and that $V \in T_C(Q)$. ∎

REMARK 1. A coordinate-free alternative for part (B) of the proof must, of course, be possible, but even for the case where $Q_i = Q$ for all i, it does not look very attractive.

REMARK 2. In the "only if" part of the proof [part (A)] we may replace n and/or m by $+\infty$. In the "if" part of the proof, however, problems may arise without further assumptions. For example, consider the case where $p = 1$, $k' = 2$, $A = 1$,

$$V = \begin{pmatrix} 1 & 1 & 1 & \cdots \\ 1 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}, \qquad Q_i = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & t_i^{1/2} & t_i^{1/3} & \cdots \\ 0 & t_i^{1/3} & t_i^{1/4} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

Then any L_2^i will contain divergent elements. On the other hand, if $p = k$, then L_2^i is empty and no problems can arise. So if $p = k$, the theorem holds for any m and any n, finite or not.

REMARK 3. If $k = \min(m, n)$, then $T_C(Q)$ is, of course, equal to the set of all m by n matrices, and if $p < k$, the theorem would be incorrect, whereas if $p = k$ it would be trivial. In view of $\operatorname{rank} Q < \operatorname{rank} Q^0$, we need only consider $k < \min(m, n)$, however.

Now that we know what $T_C(Q)$ looks like, it is an easy step to compute $N_C(Q)$, the normal cone of C at Q. Combining the definition of $N_C(Q)$ given at the beginning of this section with what we said about (Z, Q) towards the end of the previous section, we have

$$Z \in N_C(Q) \quad \text{if and only if} \quad \sum_{j,k} z_{jk} v_{jk} \leqslant 0 \quad \text{for all } V \in T_C(Q).$$

If we have that

$$Q = \begin{pmatrix} A & AR \\ LA & LAR \end{pmatrix}$$

and put

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix},$$

so that $V \in T_C(Q)$ if and only if

$$V_{22} = -LV_{11}R + LV_{12} + V_{21}R,$$

the inequality can be written as $(V_{11}, Z_{11}) + (V_{12}, Z_{12}) + (V_{21}, Z_{21}) + (V_{22}, Z_{22}) \leqslant 0$ if we put

$$Z = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{pmatrix}.$$

Now,

$$\begin{aligned}(V_{22}, Z_{22}) &= (-LV_{11}R + LV_{12} + V_{21}R, Z_{22}) \\ &= (V_{11}, -L^T Z_{22} R^T) + (V_{12}, L^T Z_{22}) + (V_{21}, Z_{22} R^T),\end{aligned}$$

and since the inequality must hold or all $V \in T_C(Q)$, it follows that

$$Z_{11} = L^T Z_{22} R^T, \qquad Z_{12} = -L^T Z_{22}, \qquad Z_{21} = -Z_{22} R^T,$$

which is equivalent to $Z(\operatorname{im} Q^T) = 0$ together with $Z(\operatorname{im}^c Q^T) = \ker Q^T$. Hence we have shown the following theorem.

THEOREM 2. *If* $\operatorname{rank} Q \leqslant k$, *with* Q *m by n, and if* $k < \min(m, n)$, *then* $Z \in N_C(Q)$ *if and only if* $Z(\operatorname{im} Q^T) = 0$ *and* $Z(\operatorname{im}^c Q^T) = \ker Q^T$.

In particular, if

$$Q = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix},$$

with A still nonsingular, then $V \in T_C(Q)$ and $Z \in N_C(Q)$ if and only if

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & 0 \end{pmatrix} \quad \text{and} \quad Z = \begin{pmatrix} 0 & 0 \\ 0 & Z_{22} \end{pmatrix}.$$

Clearly, if $\operatorname{rank} Q$ increases, the number of degrees of freedom of Z decreases, so that the inclusion $0 \in \partial f(Q) + N_C(Q)$ becomes tighter.

So far, Q has been arbitrary, but we may impose certain structural restrictions on Q. For example, we may require that Q be n by n and symmetric. Then automatically V must be symmetric, as Q and V belong to the same space, and we must also take Z symmetric, as z belongs to the conjugate space. Similarly, we may require that Q be Hankel (then V and Z too are Hankel), or that Q be reversely circulant (then V and Z too are reversely circulant).

3. COMPUTING $\partial f(Q)$ WHEN USING THE l_2-INDUCED NORM

Let Q be a finite m by n matrix, and let

$$f(Q) = |Q - Q^0|^2 = \lambda_{\max}\big((Q - Q^0)^T(Q - Q^0)\big).$$

Let us regard $f(Q)$ as a composite function

$$f = g \circ F,$$

with

$$F(Q) = M = (Q - Q^0)^T(Q - Q^0),$$

so that M is symmetric, and

$$g(M) = \lambda_{\max}(M).$$

Then we can apply a chain rule for generalized gradients; see Clarke [1, Theorem 2.3.10, Chain Rule II]. We have that F is strictly differentiable (see Clarke [1, p. 30] for a definition), and that its derivative $D_sF(Q)$ satisfies

$$D_sF(Q)(V) = (Q - Q^0)^T V + V^T(Q - Q^0) \qquad \text{for any } m \text{ by } n \text{ matrix } V.$$

Furthermore, g is Lipschitz in a neighborhood of $F(Q)$ with respect to the following norm:

$$|M| = \sup_s \{|Ms| : |s| \leqslant 1\},$$

where $|Ms|$ and $|s|$ are Euclidean norms (so that $|M|$ too is l_2-induced). For we have that

$$|M - M'| = \sup_s \{|(M - M')s| : |s| \leqslant 1\} \geqslant \sup_s \{|Ms| - |M's| : |s| \leqslant 1\}$$

$$\geqslant \sup_s \Big\{|Ms| - \sup_t \{|M't| : |t| \leqslant 1\} : |s| \leqslant 1\Big\}$$

$$= \lambda_{\max}(M) - \lambda_{\max}(M') = g(M) - g(M'),$$

and similarly, $|M - M'| \geqslant g(M') - g(M)$, so that $|g(M) - g(M')| \leqslant |M - M'|$; hence the Lipschitz constant is equal to 1. Another way of showing that g is Lipschitz in a neighborhood of $F(Q)$ is to consider the set of all symmetric n by n matrices as a subspace of R^q with $q = n^2$, and to apply Rockafellar [2, Proposition 4A], for we have that g is finite everywhere and that g is convex, because $g(M)$ is the supremum over $|s| \leqslant 1$ of linear, and hence convex, functions (Ms, s).

By Chain Rule II, mentioned above, it now follows that $\partial f(Q)$ is the set of all Z such that for some $W \in \partial g(M)$,

$$(Z, V) = (W, D_s F(Q)(V)) \qquad \text{for all } V,$$

where, as before, $(Z, V) = \sum_{i,j} z_{ij} v_{ij} = \operatorname{trace}(Z^T V)$, and similarly for $(W, D_s F(Q)(V))$. Because M is symmetric, so is W, and since $(A, B) = (A^T, B^T)$ for any A and B such that this makes sense, it follows that

$$\begin{aligned}(Z, V) &= \left(W, (Q - Q^0)^T V + V^T (Q - Q^0)\right) \\ &= \left(W, (Q - Q^0)^T V\right) + \left(W, V^T (Q - Q^0)\right) \\ &= \left(W, (Q - Q^0)^T V\right) + \left(W, (Q - Q^0)^T V\right) = 2\left(W, (Q - Q^0)^T V\right).\end{aligned}$$

Since this must be true for all V, it follows that

$$Z = 2(Q - Q^0)W, \qquad W \in \partial g(M),$$

and it remains to compute $\partial g(M)$.

Obviously, $M = (Q - Q^0)^T (Q - Q^0)$ is positive semidefinite. Further, since g is convex, $\partial g(M)$ is the set of all subgradients of g at M in the sense of convex analysis (see e.g. Rockafellar [7]), so that $W \in \partial g(M)$ if and only if

$$g(M + M') - g(M) \geqslant (W, M') \qquad \text{for all } M'.$$

If U is any orthonormal matrix, this is equivalent to

$$\begin{aligned}g(U^T M U + M') - g(U^T M U) &\geqslant (W, U M' U^T) \\ &= (U^T W U, M') \qquad \text{for all } M'.\end{aligned}$$

Choose U such that

$$U^T M U = \begin{pmatrix} \mu I_p & 0 \\ 0 & \mu I_q - D_{22} \end{pmatrix},$$

where I_p and I_q are identity matrices and $\mu = \lambda_{\max}(M)$. Further, D_{22} is a diagonal matrix with positive diagonal elements. Hence the multiplicity of the largest eigenvalue of M is p. Partition M' accordingly as

$$M' = \begin{pmatrix} M'_{11} & M'_{12} \\ M'_{21} & M'_{22} \end{pmatrix};$$

then we have that

$$g\left(\begin{pmatrix} M'_{11} & M'_{12} \\ M'_{21} & M'_{22} - D_{22} \end{pmatrix} + \mu I\right) \geqslant \mu + (U^T W U, M').$$

From the definition of g it easily follows from this that

$$g\left(\begin{pmatrix} M'_{11} & M'_{12} \\ M'_{21} & M'_{22} - D_{22} \end{pmatrix}\right) \geqslant (U^T W U, M').$$

Partition $B = U^T W U$ as

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}.$$

We will show that $B_{22} = 0$, $B_{12} = 0$, $B_{21} = 0$. First take $M'_{11} = 0$ and $M'_{12} = 0$, so that $M'_{21} = 0$, and take M'_{22} so small that $D_{22} - M'_{22}$ is positive definite. Then it follows that

$$0 \geqslant (B_{22}, M'_{22}) \qquad \text{for all } M'_{22}$$

and hence that $B_{22} = 0$. Next take $M'_{11} = 0$ and $M'_{22} = 0$, and for fixed (i, j) take the ij-element of M'_{12} equal to $m \neq 0$, and take all other elements of M'_{12} equal to zero. Then

$$g\left(\begin{pmatrix} 0 & M'_{12} \\ M'_{21} & -D_{22} \end{pmatrix}\right) \geqslant 2(B_{12}, M'_{12}) = 2m(B_{12})_{ij} = r.$$

For some diagonal element d of D_{22} we have that $r \leqslant \frac{1}{2}[-d+(d^2+4m^2)^{1/2}]$, as easily follows from the definition of g, for the eigenvalues of the matrix to the left are nonpositive, except one which is equal to $\frac{1}{2}[-d+(d^2+4m^2)^{1/2}]$. Expanding the square root gives

$$\frac{m}{d}+\cdots \geqslant 2|(B_{12})_{ij}|,$$

and since we can take m arbitrarily small, it follows that $(B_{12})_{ij}=0$, and hence that $B_{12}=0$, and also that $B_{21}=0$.

Let us now see what B_{11} looks like. Take $M'_{12}=0$, hence $M'_{21}=0$, and take $M'_{22}=0$ and

$$M'_{11}=\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_p\end{pmatrix}.$$

For some fixed i let $\lambda_i=-1$, and let $\lambda_j=0$, $j\neq i$. Then from

$$g\left(\begin{pmatrix}M'_{11} & 0\\ 0 & -D_{22}\end{pmatrix}\right)\geqslant (B_{11}, M'_{11})$$

it follows that $0\geqslant -(B_{11})_{ii}$. If we let $\lambda_j=\lambda\neq 0$ for all j, then it follows that $1\geqslant \Sigma_j(B_{11})_{jj}$ if $\lambda>0$, and $-1\geqslant -\Sigma_j(B_{11})_{jj}$ if $\lambda<0$, at least if $|\lambda|$ is small enough. Hence the diagonal elements of B_{11} are nonnegative and sum to unity. But the same must be true for $U_{11}^T B_{11}U_{11}$ for any orthonormal U_{11}, which implies that B_{11} also must be positive semidefinite. Conversely, if B_{11} is positive semidefinite and if trace $B_{11}=1$, then the last inequality above, involving M'_{11}, is satisfied for all M'_{11}, and if we take $B_{12}=0$, $B_{21}=0$, $B_{22}=0$, then $g(M+M')-g(M)\geqslant (W, M')$ for all M'. Hence we have shown the following result.

THEOREM 3. *Let M be any n by n symmetric matrix, and let $g(M)=\lambda_{\max}(M)$. Then if*

$$M=U\begin{pmatrix}\mu I_p & 0\\ 0 & *\end{pmatrix}U^T$$

with U orthonormal, and μ the largest eigenvalue of M, whose multiplicity is

equal to p, then $\partial g(M)$ *is the set of all* $W = UBU^T$ *with*

$$B = \begin{pmatrix} B_{11} & 0 \\ 0 & 0 \end{pmatrix},$$

where B_{11} *is a p by p positive semidefinite matrix such that* $\operatorname{trace} B_{11} = 1$.

REMARK 4. See Clarke [1, Example 2.8.7] for a similar result for positive definite M, which is only completely worked out if the multiplicity of $\lambda_{\max}(M)$ is equal to 1; and further see e.g. Kato [8].

REMARK 5. Another approach would be to use the fact that

$$\partial(|Q|) = \{ Z : |Z| \leqslant 1, (Z, Q) = |Q| \},$$

but this would probably require the same amount of work.

Combining this theorem with what we found earlier in this section leads to the next theorem.

THEOREM 4. *Let* Q *and* Q^0 *be any m by n matrices, and let* $f(Q) = \lambda_{\max}((Q - Q^0)^T(Q - Q^0))$. *Then* $\partial f(Q)$ *is the set of all*

$$Z = 2(Q - Q^0)W \qquad \text{with} \quad W \in \partial\lambda_{\max}(M),$$

with $M = (Q - Q^0)^T(Q - Q^0)$, *and with* W *as in Theorem* 3.

If $\partial f(Q)$ is a singleton, then f is ordinarily differentiable. This will happen if M is a positive matrix, because then the eigenvector space of $\lambda_{\max}$ is one dimensional. It is not clear, however, whether positive matrices can play an important part in the problem considered in this paper.

If we restricted ourselves to reversely circulant matrices, we could start from Theorem 4 and see what comes out. It seems better, however, to start from scratch in that case, forget about eigenvalues, and use the fact that

$$f(Q) = \sup_x \left\{ |(Q - Q^0)x|^2 : |x| = 1 \right\}.$$

Let the first row of Q and that of $Z \in \partial(|Q|^2)$ be

$$(a_1, a_2, \ldots, a_n) \quad \text{and} \quad (z_1, z_2, \ldots, z_n),$$

respectively. Then

$$\sup_x \{|Qx|^2 : |x| = 1\} = \sum a_i^2 + \sup_x \left\{ 4\Big(\sum_{i<j} a_i a_j\Big)\Big(\sum_{i<j} x_i x_j\Big) : \sum x_i^2 = 1 \right\},$$

where, of course, $x = (x_1, \ldots, x_n)$. Depending on the sign of $\sum a_i \cdot a_j$, we must compute either the supremum or the infimum of $\sum x_i x_j$ over $\sum x_i^2 = 1$, which amounts to computing the extrema of $(\sum x_i)^2 = 1 + 2\sum x_i x_j$. Since

$$\inf_x \left\{ \Big(\sum x_i\Big)^2 : \sum x_i^2 = 1 \right\} = \begin{cases} 0 & \text{if} \quad n \geqslant 2, \\ 1 & \text{if} \quad n = 1, \end{cases}$$

and

$$\sup_x \left\{ \Big(\sum x_i\Big)^2 : \sum x_i^2 = 1 \right\} = n \qquad \text{if} \quad n \geqslant 1,$$

we find

$$\sup_x \{|Qx|^2 : |x| = 1\} = \sum a_i^2 + \gamma \sum_{i<j} a_i a_j,$$

with

$$\gamma = \begin{cases} 2(n-1) & \text{if} \quad \sum_{i<j} a_i a_j \geqslant 0, \\ -2 & \text{if} \quad \sum_{i<j} a_i a_j \leqslant 0 \text{ and } n \geqslant 2, \\ 0 & \text{if} \quad \sum_{i<j} a_i a_j \leqslant 0 \text{ and } n = 1. \end{cases}$$

Let $Z \in \partial(|Q|)$; then, by definition,

$$|Q + Q'|^2 - |Q|^2 \geqslant (Z, Q') \qquad \text{for all } Q'.$$

Let the first row of Q' be $(\delta_1, \delta_2, \ldots, \delta_n)$; then this gives

$$\sum (a_i + \delta_i)^2 + \gamma' \sum_{i<j} (a_i + \delta_i)(a_j + \delta_j) - \sum a_i^2 - \gamma \sum_{i<j} a_i a_j$$

$$\geqslant n \sum_i z_i \delta_i \qquad \text{for all } \delta_i.$$

Here γ' is defined like γ, but with a_i replaced by $a_i + \delta_i$. By convexity it is sufficient to take δ_i small. If $\Sigma a_i a_j \neq 0$ and δ_i is small enough, then $\gamma' = \gamma$, and it follows that

$$nz_j = 2a_j + \gamma \sum_{i \neq j} a_i \qquad \text{for all } j,$$

so that $\partial(|Q|^2)$ is a singleton if $\Sigma a_i a_j \neq 0$.

If $\Sigma a_i a_j = 0$, we get, taking all δ_i equal to zero except δ_j for some j,

$$2a_j + \gamma' \sum_{i \neq j} a_i \geqslant nz_j \quad \text{if } \delta_j \geqslant 0, \qquad 2a_j + \gamma' \sum_{i \neq j} a_i \leqslant nz_j \quad \text{if } \delta_j \leqslant 0.$$

Hence if $\Sigma_{i \neq j} a_i \geqslant 0$, then

$$-2 \sum_{i \neq j} a_i \leqslant nz_j - 2a_j \leqslant 2(n-1) \sum_{i \neq j} a_i,$$

and if $\Sigma_{i \neq j} a_i \leqslant 0$, then

$$2(n-1) \sum_{i \neq j} a_i \leqslant nz_j - 2a_j \leqslant -2 \sum_{i \neq j} a_i,$$

at least if $n \geqslant 2$.

Conversely, if these inequalities hold with strict signs, then $Z \in \partial(|Q|^2)$, still assuming that $\Sigma a_i a_j = 0$. This completes the proof of

THEOREM 5. *If Q is reversely circulant, if its first row is $(a_1, a_2, \ldots, a_n)$, $n \geqslant 2$, and if $|Q|$ is the l_2-induced norm of Q, then*

(a) *in case $\Sigma_{i<j} a_i a_j \neq 0$, $Z \in \partial(|Q|^2)$ if and only if $nz_j - 2a_j = \gamma \Sigma_{i \neq j} a_i$ and*

(b) *in case $\Sigma_{i<j} a_i a_j = 0$, if $Z \in \partial(|Q|^2)$ then*

$$\left| nz_j - 2a_j - (n-2) \sum_{i \neq j} a_i \right| \leqslant n \left| \sum_{i \neq j} a_i \right|, \qquad i = 1, 2, \ldots,$$

and if this inequality holds with a strict sign, then $Z \in \partial(|Q|^2)$.

Here the first row of Z is $(z_1, z_2, \ldots, z_n)$.

If we want to apply this theorem to our optimization problem, we must, of course, replace Q by $Q - Q^0$.

EXAMPLE 1. Let

$$Q^0 = \begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix}$$

with rank $Q^0 = 2$; hence $\alpha^2 \neq \beta^2$. Consider

$$\inf_{Q} \left\{ |Q - Q^0|^2 : Q = \begin{pmatrix} a & b \\ b & a \end{pmatrix}, \operatorname{rank} Q \leqslant k \right\} \qquad \text{for} \quad k = 1,$$

so that $b = ra$ with $r = +1$ or $r = -1$.

If rank $Q = 0$, then $a = b = 0$ and $f(Q) = (|\alpha| + |\beta|)^2$.

If rank $Q = 1$, then $a \neq 0$ and Theorems 2 and 5 imply the existence of

$$Z = \begin{pmatrix} z_1 & z_2 \\ z_2 & z_1 \end{pmatrix} \quad \text{and} \quad z_{22}$$

such that

$$0 = \begin{pmatrix} z_1 & z_2 \\ z_2 & z_1 \end{pmatrix} + \begin{pmatrix} r^2 z_{22} & -rz_{22} \\ -rz_{22} & z_{22} \end{pmatrix} \quad \text{or} \quad z_1 = -rz_2,$$

and if $(a - \alpha)(b - \beta) \neq 0$,

$$2z_1 - 2(a - \alpha) = \gamma(b - \beta), \qquad 2z_2 - 2(b - \beta) = \gamma(a - \alpha),$$

where $\gamma = 2\operatorname{sign}[(a - \alpha)(b - \beta)]$.

(a) If $r = +1$, then $a = b$, $z_1 = -z_2$, and $(2 + \gamma)(a - \alpha) + (2 + \gamma)(a - \beta) = 0$, so that if $\gamma = -2$, we only have the condition that $(a - \alpha)(a - \beta) < 0$, or

$$\min(\alpha, \beta) \leqslant a = b \leqslant \max(\alpha, \beta),$$

where the equalities result from $(a - \alpha)(a - \beta) = 0$; whereas if $\gamma = +2$ we obtain $2a = \alpha + \beta$, and $2 = 2\operatorname{sign}[-(\alpha - \beta)^2]$, a contradiction.

(b) If $r = -1$, a similar analysis shows that

$$\min(\alpha, -\beta) \leqslant a = -b \leqslant \max(\alpha, -\beta).$$

It follows that $f_{\min} = (|\alpha| - |\beta|)^2$ and that for optimal Q we have that $\operatorname{rank} Q = 1$, unless either $\alpha = 0$ or $\beta = 0$, although even then we can rely on there being some Q with $\operatorname{rank} Q = 1$. Notice that there is no unique optimal Q.

4. COMPUTING $\partial f(Q)$ WHEN USING THE FROBENIUS NORM

If instead of the l_2-induced norm we use the Frobenius norm, we obtain a much simpler result. Then

$$f(Q) = |Q - Q^0|^2 = \sum_{i,j} \left(q_{ij} - q_{ij}^0\right)^2,$$

and it easily follows that

$$\partial f(Q) = \{2(Q - Q^0)\},$$

which for all Q is a singleton (the same is true for all $Q \neq Q^0$ if $f(Q) = |Q - Q^0|$).

EXAMPLE 2. Let Q^0, Q, and k be as in Example 1. If $Q = 0$ then $f(Q) = 2(\alpha^2 + \beta^2)$. If $Q \neq 0$, then z_{22} must exist such that

$$0 = 2\begin{pmatrix} a-\alpha & b-\beta \\ b-\beta & a-\alpha \end{pmatrix} + \begin{pmatrix} r^2 z_{22} & -r z_{22} \\ -r z_{22} & z_{22} \end{pmatrix}, \qquad \text{where} \quad r^2 = 1 \text{ and } b = ra.$$

It follows that $a - \alpha = -r(b - \beta)$. If $r = +1$ then $a = b = \frac{1}{2}(\alpha + \beta)$, and if $r = -1$ then $a = -b = \frac{1}{2}(\alpha - \beta)$, giving $f(Q) = (\alpha - \beta)^2$ and $f(Q) = (\alpha + \beta)^2$, respectively, and $f_{\min} = (|\alpha| - |\beta|)^2$. Notice that the optimal solution is unique.

5. RELATING THE STRUCTURE OF A REVERSELY CIRCULANT MATRIX TO ITS RANK

From now on we restrict our attention to reversely circulant matrices. First of all we note that if the rank of a reversely circulant matrix is equal to k, then its submatrix formed from the first k rows and the first k columns is nonsingular, so that such a matrix has the form as indicated in Theorem 6 below. In this theorem we give a necessary and sufficient condition for a matrix of this form to be reversely circulant. This can be used to construct reversely circulant matrices given their rank.

THEOREM 6. *Let the n by n matrix Q have the form*

$$Q = \begin{pmatrix} A & AR \\ R^T A & R^T AR \end{pmatrix}, \qquad \text{with } A \text{ } p \text{ by } p,$$

and let the first row of R^T be $(r_1, \ldots, r_p)$. Consider the circulant matrix

$$C = \begin{pmatrix} r_1 & r_2 & \cdots & r_p & -1 & 0 & \cdots & 0 \\ 0 & r_1 & \cdots & r_{p-1} & r_p & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ r_2 & r_3 & \cdots & -1 & 0 & 0 & \cdots & r_1 \end{pmatrix}.$$

If Q is reversely circulant and if $\operatorname{rank} Q = p$, *then* $\operatorname{rank} C = n - p$. *Conversely, if C has the indicated form and if* $\operatorname{rank} C = n - p$, *then if we let $(a_1, \ldots, a_n)$ be the first row of Q, and if we prescribe $(a_1, \ldots, a_p)$ and compute $(a_{p+1}, \ldots, a_n)$ according to*

$$\begin{aligned} a_{p+1} &= r_1 a_1 + \cdots + r_p a_p, \\ a_{p+2} &= r_1 a_2 + \cdots + r_p a_{p+1}, \\ &\vdots \\ a_n &= r_1 a_{n-p} + \cdots + r_p a_{n-1}, \end{aligned} \qquad (*)$$

then $(a_1, \ldots, a_n)$ determines a reversely circulant matrix Q, and $\operatorname{rank} Q \leqslant p$.

Proof. If Q has the indicated form, and if Q is reversely circulant, then, if $(a_1, \ldots, a_n)$ is the first row of Q, we have, of course, that $(*)$ holds, and hence that $QC = 0$. If in addition $\operatorname{rank} Q = p$, it follows from this that

rank $C \leqslant n-p$, but C contains an $n-p$ by $n-p$ triangular matrix with -1 everywhere on the main diagonal, so that rank $C \geqslant n-p$; hence rank $C = n-p$. Conversely, if we prescribe $(a_1,\ldots,a_p)$ and compute $(a_{p+1},\ldots,a_n)$ from $(*)$ given $r_1,\ldots,r_p$, we get $QC=0$ if Q is the reversely circulant matrix generated by $(a_1,\ldots,a_n)$. If rank $C = n-p$, it follows that rank $Q \leqslant p$. ■

EXAMPLE 3. Let $n=4$ and rank Q = rank $A = p = 2$. Then

$$Q = \begin{pmatrix} a & b & \pm a & \pm b \\ b & \pm a & \pm b & a \\ \pm a & \pm b & a & b \\ \pm b & a & b & \pm a \end{pmatrix} \quad \text{with} \quad r_1^2 = 1 \text{ and } r_2 = 0,$$

and

$$R^T = \begin{pmatrix} r_1 & 0 \\ 0 & r_1 \end{pmatrix}.$$

If we let rank Q = rank $A = p = 1$, then

$$Q = \begin{pmatrix} a & \pm a & a & \pm a \\ \pm a & a & \pm a & a \\ a & \pm a & a & \pm a \\ \pm a & a & \pm a & a \end{pmatrix} \quad \text{with} \quad r_1^2 = 1$$

and

$$R = (r_1, 1, r_1).$$

The numbers $r_1,\ldots,r_p$ can be computed from p nonlinear equations as follows. If rank $C = n-p$, since C is circulant, there must exist numbers $s_1,\ldots,s_{n-p}$ such that

$$\begin{aligned} &(-1,0,\ldots,0,r_1,r_2,\ldots,r_p) \\ &\quad = s_1(r_1,r_2,\ldots,r_p,-1,0,\ldots,0) \\ &\qquad + s_2(0,r_1,\ldots,r_{p-1},r_p,-1,\ldots,0) \\ &\qquad + \cdots \\ &\qquad + s_{n-p}(0,0,\ldots,r_{2p-n+1},r_{2p-n+2},r_{2p-n+3},\ldots,-1), \end{aligned}$$

where $r_i = 0$ if $i \leqslant 0$, because the left hand side of this equation is the

$(n-p+1)$st row of C. The s_j can be eliminated:

$$\begin{aligned} r_p &= -s_{n-p}, \\ r_{p-1} &= -s_{n-p-1} + s_{n-p}r_p, \\ r_{p-2} &= -s_{n-p-2} + s_{n-p-1}r_{p-1} + s_{n-p}r_p, \\ &\vdots \end{aligned}$$

and then p equations in $r_1, \ldots, r_p$ result.

EXAMPLE 4. $n = 5$ and $p = 3$ gives

$$(-1, 0, r_1, r_2, r_3) = s_1(r_1, r_2, r_3, -1, 0) + s_2(0, r_1, r_2, r_3, -1),$$

so that $s_2 = -r_3$, $s_1 = -r_2 - r_3^2$, and the resulting equations are $r_1(r_2 + r_3^2) = 1$, $r_1r_3 + r_2^2 + r_2r_3^2 = 0$, $r_1 + 2r_2r_3 + r_3^3 = 0$, giving $r_1 = 1$, $r_2 = -r_3 = \frac{1}{2}(-1 \pm \sqrt{5})$. If $p = n-1$, everything is very simple: $(-1, r_1, \ldots, r_{n-1}) = s_1(r_1, r_2, \ldots, -1)$, giving $r_i = (-1)^{i+1}r_1^i$, $i = 1, \ldots, n-1$ and $r_1r_{n-1} = 1$, so that $(-1)^n r_1^n = 1$. If n is even, this leads to $r_1^2 = 1$ and hence to $(r_1, \ldots, r_{n-1}) = (r_1, -1, r_1, -1, \ldots)$, and if n is odd, it leads to $r_1 = -1$ and hence to $r_i = -1$ for all i.

Since R depends only on $r_1, \ldots, r_p$, R can be determined once we know the r_i. This could be done as follows. From $QC = 0$ we have that $(A \; AR)C = 0$, or, if A is nonsingular, $(I \; R)C = 0$. Partitioning C as

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

in such a way that C_{11} is p by p, so that C_{22} is nonsingular, we obtain

$$R = -C_{12}C_{22}^{-1}.$$

Obviously, we may also first permute the columns of C before partitioning C. In particular, if we apply a cyclic permutation in such a way that C_{22} becomes a lower triangular matrix with -1 everywhere on the main diagonal, this may lead to a relatively easy computation, as a triangular matrix is easily inverted. (Actually, we may also put $R = -C_{11}C_{12}^{-1}$ if we let C_{11} be $n-p$ by p, and if we do not permut C at all.)

Apart from $R = -C_{12}C_{22}^{-1}$, it follows from $(I\ R)C = 0$ that $C_{11} = C_{12}C_{22}^{-1}C_{21}$, which gives a (redundant) system of equations for $r_1, \ldots, r_p$.

6. SOLVING THE MINIMIZATION PROBLEM IN THE CASE OF REVERSELY CIRCULANT MATRICES AND THE FROBENIUS NORM

In order to solve for $\inf_Q\{|Q - Q^0|^2 : \operatorname{rank} Q \leqslant k\}$, $k < \operatorname{rank} Q^0$, with Q^0 and Q reversely circulant, and where $|Q - Q^0|$ is the Frobenius norm of $Q - Q^0$, we could let $p = k, k-1, \ldots,$ and for each of these p solve for a p by p nonsingular matrix A, an $n-p$ by $n-p$ matrix Y, and a p by $n-p$ matrix R from

$$2(Q - Q^0) = 2\begin{pmatrix} A & AR \\ R^TA & R^TAR \end{pmatrix} - 2Q^0 = \begin{pmatrix} RYR^T & -RY \\ -YR^T & Y \end{pmatrix},$$

as follows from Theorem 2 and Section 4. Let

$$Q^0 = \begin{pmatrix} Q_{11}^0 & Q_{12}^0 \\ Q_{21}^0 & Q_{22}^0 \end{pmatrix};$$

then this is easily seen to be equivalent to

$$A = (Q_{11}^0 + Q_{12}^0R^T)(I + RR^T)^{-1},$$

$$Y = R^TAR - Q_{22}^0,$$

$$(I + RR^T)AR = Q_{12}^0 + RQ_{22}^0,$$

$$R^TA(I + RR^T) = Q_{21}^0 + Q_{22}^0R^T.$$

Note that setting $\operatorname{rank} Q = p$, with p fixed, means that $T_C(Q)$ becomes Bouligand's contingent cone of C at Q (see end of Section 1).

EXAMPLE 5. Let $n = 4$, $k = 3$, and

$$Q^0 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

If rank $A = p = 3$ then $R^T = (r_1, -1, r_1)$, where $r_1^2 = 1$ (see the case $p = n - 1$ considered below Example 4). Hence

$$A = \frac{1}{4}\begin{pmatrix} r_1 & -1 & r_1 \\ -1 & r_1 & 3 \\ r_1 & 3 & r_1 \end{pmatrix} \quad \text{and} \quad Y = \tfrac{1}{4}r_1,$$

so that $|Q - Q^0| = 1$. If rank $A = p = 2$, then

$$R^T = r_1\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad r_1^2 = 1$$

and

$$A = \tfrac{1}{2}r_1\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad Y = \tfrac{1}{2}r_1\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix};$$

hence $|Q - Q^0| = \sqrt{2}$. Finally, if rank $A = 1$, then $R = (r_1, 1, r_1)$, $r_1^2 = 1$, and

$$A = \tfrac{1}{4}r_1, \qquad Y = \frac{1}{4}\begin{pmatrix} r_1 & -3 & r_1 \\ -3 & r_1 & 1 \\ r_1 & 1 & r_1 \end{pmatrix},$$

so that $|Q - Q^0| = \sqrt{3}$.

It follows that $f_{\min} = 1$, which happens to be equal to the minimum if any Q, reversely circulant or not, is allowed. The same is true if instead of $k = 3$ we take $k = 2$ or $k = 1$. Note that Q^0 is itself reversely circulant.

OPEN QUESTION. It is an open question if, when Q^0 is reversely circulant, there always exists a reversely circulant Q for which the value of $f_{\min}$ is assumed. If Q^0 is just Hankel (and finite), then there may be no Q that is Hankel for which $f_{\min}$ is assumed, at least if one uses the l_2-induced norm.

COUNTEREXAMPLE (Heij [9]). Take $n = 3$ and let

$$Q^0 = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

Then if the norm is the l_2-induced norm, no Q exists which is Hankel, and for which the minimum of $|Q - Q^0|$ is assumed, under the condition that $\operatorname{rank} Q \leqslant 1$.

Another interesting fact is that the minimum may be assumed for a Q with $\operatorname{rank} Q < k$, an occurrence that apparently is rare when considering infinite Hankel matrices and again the l_2-induced norm.

EXAMPLE 6. Let $n = 3$ and $k = 2$, and take

$$Q^0 = \begin{pmatrix} \alpha & \beta & \gamma \\ \beta & \gamma & \alpha \\ \gamma & \alpha & \beta \end{pmatrix}.$$

Then $f_{\min} = \min\{(\alpha+\beta+\gamma)^2; (\alpha-\beta)^2+(\beta-\gamma)^2+(\gamma-\alpha)^2\}$, where $(\alpha+\beta+\gamma)^2$ is obtained if we let $\operatorname{rank} Q = 2$, and $(\alpha-\beta)^2+(\beta-\gamma)^2+(\gamma-\alpha)^2$ if $\operatorname{rank} Q = 1$. Depending on the values of α, β, and γ, it may be better to let $\operatorname{rank} Q = 1$. That Q with $\operatorname{rank} Q = 1$ might be preferred can be illustrated geometrically. If we let

$$Q = \begin{pmatrix} a & b & c \\ b & c & a \\ c & a & b \end{pmatrix},$$

then because $\operatorname{rank} Q \leqslant 2$, we have that $a^3 + b^3 + c^3 = 3abc$ or $(a+b+c)^3 = 3(a+b+c)(ab+bc+ca)$. If $a+b+c = 0$, then $\operatorname{rank} Q$ is equal to 2 or 0, and if $(a+b+c)^2 = 3(ab+bc+ca)$ or $(a-b)^2+(b-c)^2+(c-a)^2 = 0$, then $\operatorname{rank} Q$ is equal to 1 or 0. So the Q's with rank equal to 2 or 0 form a 2-dimensional subspace in R^3, and those with rank equal to 1 or 0 form a 1-dimensional subspace in R^3. These two subspaces meet in 0 only, and depending on whether (α, β, γ) is closer to the latter than to the former, we must let $\operatorname{rank} Q = 1$ or $\operatorname{rank} Q = 2$.

Thanks are due to C. Heij and J. W. Nieuwenhuis for their contributions.

REFERENCES

1 F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, 1983.
2 R. T. Rockafellar, The theory of subgradients and its applications to problems of optimization. Convex and Nonconvex Functions, Heldermann, 1981.

3 R. T. Rockafellar, Generalized directional derivatives and subgradients of nonconvex functions, *Canad. J. Math.* XXXII:257–280 (1980).
4 G. W. Stewart, *Introduction to Matrix Computations*, Academic, 1973.
5 K. Glover, All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds, *Internat. J. Control* 39:1115–1193 (1984).
6 J. W. Nieuwenhuis, private communication.
7 R. T. Rockafellar, *Convex Analysis*, Princeton, 1970.
8 T. Kato, *Perturbation Theory for Linear Operators*, Springer, 1966.
9 C. Heij, Private communication.

Received 30 July 1985; revised 18 June 1986